GraphVid: Interactive Graph-Controllable Video Generation

Vedant Shah[1], Onkar Susladkar[1], Tushar Prakash[2], Kiet A. Nguyen[1], Tianjiao Yu[1], Adheesh Juvekar[1], Muntasir Wahed[1], and Ismini Lourentzou[1]

[1]University of Illinois Urbana-Champaign [2]Sony Research India

{vrshah4,onkarks2,lourent2}@illinois.edu

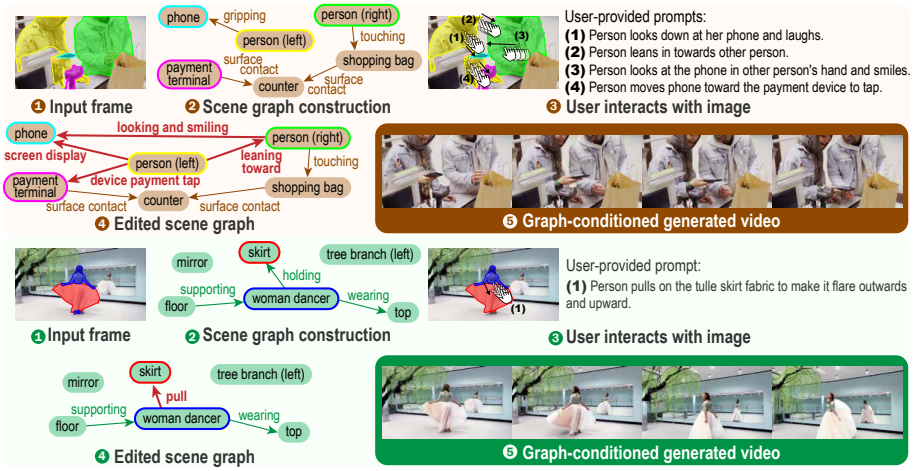


Fig. 1: GraphVid enables controllable multi-object image-to-video generation through interaction graphs. Starting from an input frame, GraphVid constructs a scene graph that represents entities and their interactions. Users can directly modify these interactions by interacting with the image to specify desired dynamics. The edited interaction graph is then translated into conditioning tokens that guide a frozen video diffusion backbone to generate a coherent video sequence.

Abstract. Controllable video generation remains challenging due to the difficulty of specifying precise multi-object interactions using text prompts or motion-control inputs that primarily constrain pixel movement. In practice, trajectory-based control often requires users to draw accurate tracks for multiple objects, which scales poorly with scene complexity and becomes ambiguous under occlusion or overlap. To enable flexible yet precise multi-subject control, we introduce **GraphVid**, a graph-conditioned image-to-video generation model that enables interactive control through structured interaction graphs. We further curate GraphVid-Bench, a large-scale interaction-centric video dataset with structured relational annotations to enable training of interaction-aware video generation models. Despite using substantially less training data and fewer trainable parameters than prior motion-control methods, GraphVid delivers strong controllability and video quality. Compared with Motion-I2V, GraphVid reduces FID by up to 39.9% and FVD by 37.6%, while improving PSNR (9.87→15.98) and SSIM (0.38→0.61). Our

results highlight the potential of structured semantic interfaces as a powerful paradigm for controllable video generation.
PLAN Lab https://plan-lab.github.io/graphvid

Keywords: Controllable Video Generation · Scene Graphs

1 Introduction

Image-to-Video (I2V) generation has attracted increasing attention in generative modeling [5, 34, 42, 43, 45]. Given a single image, the goal is to synthesize a temporally coherent video that preserves visual identity while producing realistic motion. This task is inherently challenging since the model must infer plausible temporal dynamics from a single static frame, where motion, interactions, and physical evolution are fundamentally ambiguous. As a result, generating videos that are both temporally consistent and physically plausible remains difficult.

Recent work on controllable I2V generation has largely focused on trajectory-conditioned motion control [8,29,42,53]. Methods such as Motion Prompting [10], Wan-Move [8], MagicMotion [32], and Motion-I2V [42] guide diffusion-based video models using explicit motion signals, including point trajectories, optical flow, segmentation masks, or sparse bounding boxes. These signals are transformed into dense motion representations and injected into the generative backbone through mechanisms such as ControlNet [58] adapters, latent feature manipulation, or motion-aware attention, enabling user-specified object and camera motion while preserving appearance. Among these, Wan-Move achieves strong performance by directly modifying latent conditioning features of a large-scale Diffusion Transformer (DiT) [37] model and training on millions of videos, improving motion fidelity and generalization. However, these methods treat motion primarily as geometric displacement, overlooking the physical causes governing object dynamics. To address this, a few physics-aware approaches introduce control signals based on forces or velocity fields to model object interactions [12,39], jointly model visual and latent physical dynamics [41], or incorporate structured physical knowledge by injecting textual descriptions, qualitative categories, and quantitative physical properties through physics-aware attention [49].

Despite recent progress, existing approaches for controllable video generation remain limited in how they model scene dynamics: **(1)** Current physics-aware approaches introduce control signals based on forces, velocities, or structured physical annotations [12,39,49]. However, these methods typically encode physics as low-level control fields or predefined categories, which limits their ability to represent complex interactions between multiple entities in open-domain scenes. **(2)** Methods that condition on low-level geometric cues can make control cumbersome in multi-object scenarios, as users must author accurate signals for each object, and small errors in a few controls can yield globally implausible motion (*e.g.*, inconsistent contact, sliding, penetrations, or desynchronized behavior). **(3)** Moreover, both trajectory-based and physics-conditioned methods rely heavily on large-scale curated datasets containing dense motion annotations, trajectories, optical flows, or physics labels, which are expensive to obtain

and difficult to scale across diverse environments. For instance, Wan-Move [8] is trained on 2M samples, Motion Prompting [10] uses 2.2M, and Motion-I2V [42] is trained on 10M samples. This dependence restricts the practicality and generalization of existing controllable video generation pipelines.

To address these limitations and enable flexible interactive multi-object control, we introduce **GraphVid**, a graph-conditioned image-to-video generation framework that represents motion control through directed interaction scene graphs. Starting from an input image, GraphVid constructs a scene graph over detected entities with relational edges describing interactions (Figure 1). Users interact directly with the image to specify desired dynamics, and the edited interaction graph is then translated into conditioning tokens that guide a frozen video diffusion backbone to generate a coherent video sequence. To model interaction-driven dynamics, we introduce Edge-Aware Graph Reasoning, which conditions message passing on directed edge attributes so that object representations explicitly reflect their relational roles before being mapped to diffusion conditioning. By injecting this representation into a pretrained video DiT model and using LoRA modules in attention layers, the model learns to animate the scene using its own generative priors while respecting the specified interaction structure, enabling a more intuitive and interactive mechanism for motion control. To support training, we construct GraphVid-Bench, an interaction-centric dataset of ∼27K video clips paired with explicit interaction graphs. Despite using ∼24× fewer parameters and substantially less training data, GraphVid achieves competitive performance with state-of-the-art motion-control video generation methods, significantly improving perceptual quality and reconstruction fidelity, and reducing FID by up to 39.9% and FVD by 37.6% compared to Motion-I2V, while improving PSNR (9.87→15.98) and SSIM (0.38→0.61). Our contributions are:

- We propose **GraphVid**, to the best of our knowledge, the first framework that enables interactive control of multi-object dynamics through directed interaction scene graphs constructed from a single image. GraphVid's graph-to-token conditioning interface injects interaction-aware representations into a frozen video diffusion transformer using parameter-efficient LoRA adaptation, enabling structured control without degrading pretrained generative priors.
- We introduce an **Edge-Aware Graph Reasoning module** that conditions graph message passing on directed edge attributes, allowing model representations to capture complex multi-object relations and interaction-driven motion propagation across multiple entities.
- We curate **GraphVid-Bench**, a high-quality interaction-centric video generation dataset of ∼27K video clips paired with structured interaction graphs to support training and evaluation of relational video control.
- Despite requiring only 0.6B trainable parameters and achieving the fastest inference time among competing methods, GraphVid delivers substantial quality gains across perceptual and reconstruction metrics, reducing FID by 34.5% (25.98 → 17.02) and FVD by 7.8% (107.89 → 99.42), while increasing PSNR by 5.9% (15.08 → 15.98) and SSIM by 15.0% (0.53 → 0.61) compared to WISA (physics-centric text inputs), and outperforming trajectory-based baselines

such as Tora (FID −30.3%, FVD −10.0%, PSNR +41.7%, SSIM +12.9%) and Motion-I2V (FID −39.9%, FVD −37.6%, PSNR +61.9%, SSIM +60.5%).

2 Related Work

Video Generation. Recent advances in generative modeling have significantly improved the quality and scale of video generation systems [4,48,56]. Early approaches extended image diffusion models to the temporal domain using U-Net backbones and temporal modules to capture motion dynamics [14,15]. Subsequent work introduced latent video diffusion and motion-conditioned architectures to improve temporal consistency and scalability [22,55]. More recently, large-scale diffusion transformers (DiTs) have emerged as the dominant architecture for high-quality video synthesis. Models such as CogVideoX [56] and Wan [48] demonstrate that scaling transformer-based diffusion models enables long-range temporal coherence and improved visual fidelity. Recent works start from such strong pretrained video backbones and add a small number of trainable parameters that inject the desired control signal while preserving the backbone's generative priors [8,31,42,52]. This trend reflects both compute constraints and the empirical observation that large-scale backbones already encode rich appearance and temporal dynamics [10,12,29,40,44]. GraphVid follows this paradigm by keeping the video backbone frozen and learning lightweight modules that map structured interaction cues into conditioning tokens for generation.

Controllable Video Generation. While general video generation models can synthesize realistic videos, controlling the generated content remains a major challenge. Early controllable generation approaches introduced conditioning signals such as pose, depth, or motion trajectories to guide video synthesis [10,29,42]. Other works explored editing-based pipelines that modify existing videos while preserving temporal consistency [11,28,33]. More recent methods focus on integrating structured conditioning signals or intermediate representations to enable more precise control. For example, models can condition on spatial layouts [18,50], motion fields [29,42], or object trajectories [53,57] to guide generation while maintaining coherence with pretrained video backbones. Despite their strong performance, trajectory-based methods rely heavily on dense point-tracking supervision [24], making them data-intensive and difficult to scale to higher-level relational reasoning and complex interactions between objects. In contrast, we propose GraphVid, which represents interactions as a directed scene graph and explicitly models multi-object control using an edge-aware GNN to generate more plausible and physically consistent videos.

Scene Graphs for Visual Generation. Scene graphs provide a structured representation of objects and their relations within a scene and have been widely studied for image understanding and generation [7,9,25,30], 3D scene understanding [1,36,47], compositional and commonsense reasoning [20,26], *etc.*, enabling models to reason about compositional structure and object interactions. Recent graph-to-video formulations support compositional generation from object-centric interaction graphs [3]. In contrast, GraphVid targets *interactive* control

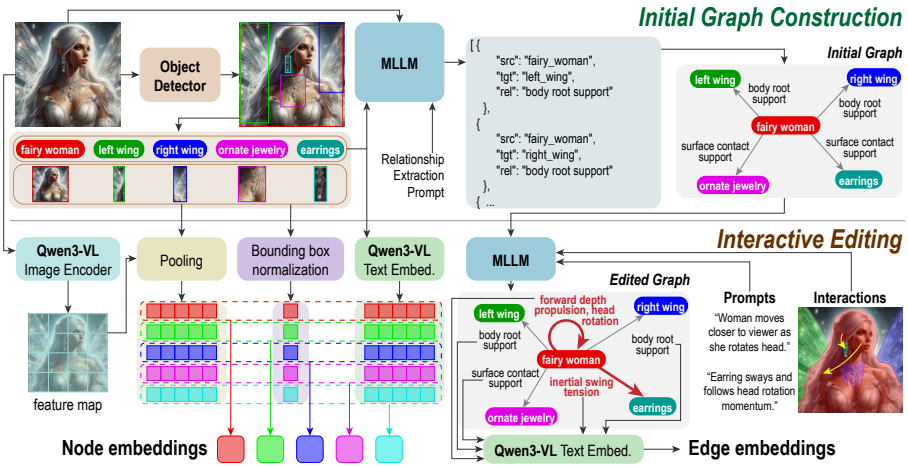


Fig. 2: GraphVid Overview. From an input image, objects are detected and encoded into node embeddings, while an MLLM extracts relational edges to construct an initial interaction graph. Edge-aware graph reasoning converts node and edge representations into conditioning tokens that guide a frozen video diffusion transformer to generate interaction-consistent video dynamics.

with an *interaction-grounded* graph and learns lightweight modules that translate graph cues into conditioning tokens compatible with a video generation backbone, enabling structured control while retaining the backbone's strong priors.

3 Method

We present **GraphVid**, a framework for controllable image-to-video (I2V) generation from a single conditioning image. The key idea is to represent user intent as a *structured interaction graph* over entities in the scene, and to interface this graph with a large pretrained video diffusion transformer using a lightweight graph-to-token adapter and LoRA-based conditioning. By keeping the backbone frozen and learning only small modules, GraphVid introduces structured, object-centric control while preserving the strong generative priors of the pretrained model. Figure 2 presents an overview.

Given a conditioning image I_0, our goal is to generate a video sequence $\{I_t\}_{t=1}^{T}$ that follows a user-specified set of object interactions. To express scene dynamics compositionally, we construct an attributed directed interaction graph $G = (V, E)$, where each node $v_i \in V$ corresponds to an entity detected in I_0 and each directed edge $e_{ij} \in E$ encodes a relation type $r_{ij} \in \mathcal{R}$ from object i to object j (*e.g.*, *push*, *pull*, *lift*, *hold*), describing how entities influence each other, and optional qualitative physical cues (*e.g.* coarse direction and magnitude), allowing users to describe scene dynamics in intuitive relational terms. Unary directives that involve a single entity are represented as self-loop edges e_{ii} (*e.g.*, `rotate`, `move`, `scale`). We aim to learn a conditional video generator that models $p(\mathbf{I}_{1:T} \mid I_0, G)$ where $\mathbf{I}_{1:T} = \{I_t\}_{t=1}^{T}$ denotes the T generated frames.

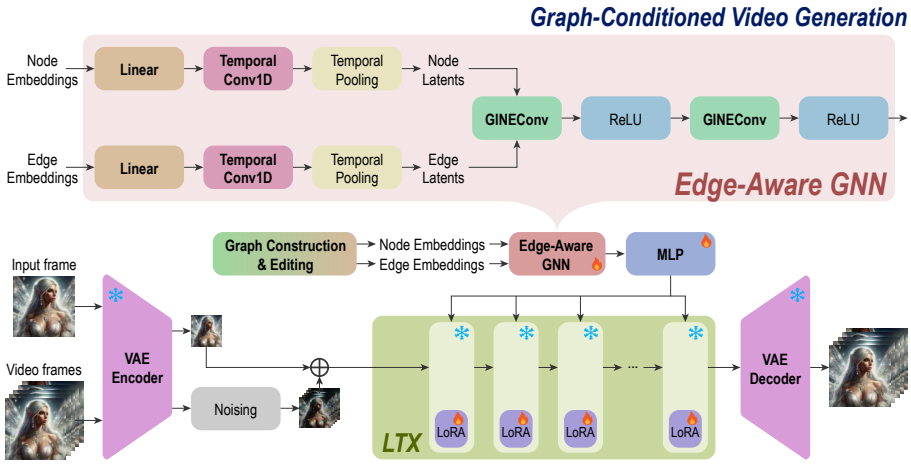


Fig. 3: GraphVid Edge-Aware Graph Reasoning. Node and edge embeddings are encoded into temporal latents and processed by our proposed edge-aware GNN that injects interaction semantics directly into message passing. The resulting graph embeddings capture relational dynamics between entities and are mapped into conditioning tokens that modulate a frozen diffusion video backbone via LoRA adapters.

Node Representations. We extract object entities from the input frame I_0 using a pre-trained vision-language detector. For each detected object v_i, we construct a unified feature vector by fusing: (i) a visual embedding $\mathbf{f}_i^{vis} \in \mathbb{R}^{d_v}$ extracted by encoding the region of I_0 inside the object's bounding box $\mathbf{b}_i$ and pooling the resulting features, and (ii) a text embedding $\mathbf{f}_i^{txt} \in \mathbb{R}^{d_t}$ encoding the object label. Normalized bounding box coordinates $\mathbf{b}_i \in \mathbb{R}^4$ are concatenated with these embeddings to form a representation $\mathbf{x}_i = [\mathbf{f}_i^{vis} \,\|\, \mathbf{f}_i^{txt} \,\|\, \mathbf{b}_i]$ that jointly captures object appearance, semantic identity, and spatial grounding.

Edge Representations. Each directed edge e_{ij} is represented with an open-vocabulary textual descriptor of the intended interaction between the source and target entities. Edges are encoded using a text embedding model to obtain the edge feature $\mathbf{e}_{ij} \in \mathbb{R}^{d_e}$.

3.1 Edge-Aware Graph Reasoning

Standard GCNs treat edges as binary connectivity signals and primarily propagate node features across the graph [17, 27]. However, in physical interaction modeling, the type of relationship between objects directly determines their motion dynamics and implies fundamentally different motion patterns and causal effects. Ignoring these relational attributes reduces the graph to simple spatial adjacency and prevents the model from reasoning about how interactions influence object behavior. To address this, we employ an edge-aware Graph Isomorphism Network (GINEConv [54]) that explicitly incorporates edge semantics during message passing. As illustrated in Figure 3, node and edge features are first projected into a shared hidden space $\mathbf{h}_i^{(0)} = \phi_n(\mathbf{x}_i), \quad \mathbf{a}_{ij} = \phi_e(\mathbf{e}_{ij})$, where

ϕ_n and ϕ_e are learnable MLPs that map node/edge features to a common embedding space $\mathbb{R}^{d_h}$. We then apply L GINEConv layers to propagate relational information across the graph

$$\mathbf{h}_i^{(l+1)} = \mathrm{MLP}^{(l)}\left(\mathbf{h}_i^{(l)} + \sum_{j \in \mathcal{N}(i)} \mathrm{ReLU}(\mathbf{h}_j^{(l)} + \mathbf{a}_{ji})\right). \tag{1}$$

Unlike standard GCNs, the GINE formulation injects edge attributes directly into the aggregation step, allowing interaction semantics to influence node updates. Consequently, each node representation is conditioned not only on neighboring objects but also on the physical relationships governing their interactions. After L layers, we obtain interaction-aware node embeddings $\mathbf{h}_i \in \mathbb{R}^{d_h}$ that encode both object identity and its dynamic relational state within the scene. These embeddings provide the downstream generative model with a structured representation of which entities should move and how they interact, enabling more physically coherent video generation.

3.2 Graph-to-Video Adapter

Modern Diffusion Transformers (DiTs) trained on large-scale video datasets provide strong spatio-temporal priors for video generation [35, 56]. Training such models from scratch is computationally prohibitive and may compromise the rich generative capabilities learned from large-scale data. Therefore, we leverage a pretrained DiT backbone [16] and keep it frozen, introducing a lightweight adapter that converts graph embeddings into conditioning tokens compatible with the transformer's input space.

Each interaction-aware node embedding $\mathbf{h}_i$ is therefore projected into the transformer latent space as $\mathbf{z}_i = \mathrm{MLP}(\mathbf{h}_i) \in \mathbb{R}^{d_l}$, where d_l denotes the hidden size of the pretrained DiT. Each $\mathbf{z}_i$ is treated as a semantic conditioning token. Since scenes contain varying numbers of objects, node sequences are padded to a maximum length $N_{\max}$. The resulting tokens, concatenated with edge text tokens, are passed as encoder hidden states. This enables the transformer's self-attention to reason over entity interactions and translate graph structure into motion during generation.

3.3 Training Objective

Since the pretrained DiT backbone is kept frozen to preserve its strong generative priors, the model requires a lightweight mechanism to adapt to the new graph conditioning modality. We therefore inject Low-Rank Adaptation (LoRA) [21] modules into the $\mathbf{Q}$, $\mathbf{K}$, $\mathbf{V}$, and output projection layers of each transformer block. Modulating these attention projections allows the model to incorporate graph tokens into the attention computation while introducing only a small number of trainable parameters. This enables the transformer to attend to interaction-aware graph embeddings without modifying the pretrained weights.

We train GraphVid using the flow-matching objective employed by the pretrained diffusion model. Given a target video $\mathbf{x}_0$, we encode it into the latent space and sample a timestep $t \sim p(t)$ from a logit-normal distribution. Noise is then added to obtain the noisy latent $\mathbf{x}_t$. The model is optimized using loss

$$\mathcal{L}_{\mathrm{CFM}}(\theta) = \mathbb{E}_{\mathbf{c}\sim\mathcal{I}_0,G}\,\mathbb{E}_{t\sim p(t)}\,\mathbb{E}_{\mathbf{x}_0\sim p_{\mathrm{data}}(\cdot|\mathbf{c}),\ \mathbf{x}_1\sim\mathcal{N}(0,\mathbf{I})}\Big[\big\|\mathbf{v}_\theta(\mathbf{x}_t,t,\mathbf{c}) - (\mathbf{x}_1-\mathbf{x}_0)\big\|_2^2\Big], \tag{2}$$

where $\mathbf{x}_t = (1-t)\mathbf{x}_0 + t\mathbf{x}_1$. During training, only the GNN parameters, Graph-to-Adapter MLP layers, and LoRA modules are updated (as θ), while the VAE and DiT backbones remain frozen. This design enables end-to-end learning of a differentiable interactive/symbolic-to-latent mapping from interaction graphs to video dynamics while retaining the pretrained model's generative capabilities.

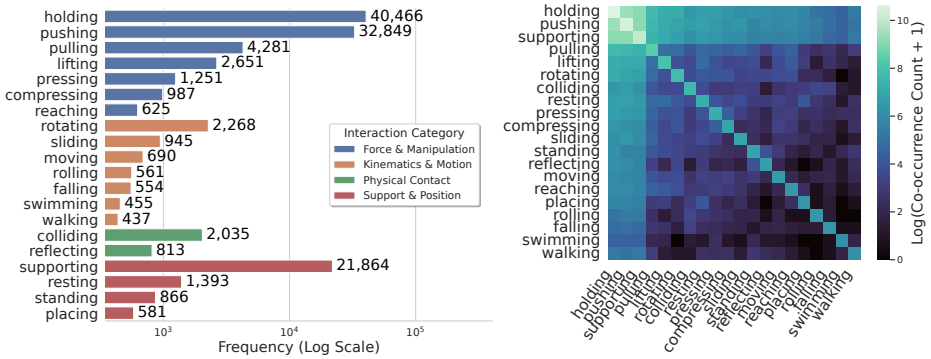
4 GraphVid-Bench Dataset

Current video generation benchmarks primarily rely on text prompts [23] or low-level motion cues [8], which provide limited control over object-level interactions. However, many real-world dynamics are fundamentally relational, where actions emerge from how entities interact with one another (*e.g.*, a person turning their head, an object moving relative to its surroundings). To enable structured control over such dynamics, we introduce GraphVid-Bench, a dataset that pairs videos with interaction graphs describing entities and their directed interaction cues.

GraphVid-Bench contains approximately **27K** interaction-centric video clips. To ensure consistent training and evaluation, all clips are standardized to 81 frames at 16 fps. Each video is paired with a directed interaction graph where nodes represent scene entities and edges encode interaction cues between them. Node attributes include visual features, semantic labels, and spatial information, while edges capture the type and direction of interactions that influence scene dynamics. As shown in Table 1, the dataset exhibits diverse structural complexity, with an average of 7.76 nodes and 4.85 edges per graph. In total, 10,982 samples contain a single interaction, and 12,641 contain multiple interactions ($N > 1$). As illustrated in Figure 4, GraphVid-Bench covers a broad spectrum of canonical physical interactions. The co-occurrence statistics further reveal that many interactions appear jointly within the same clips, indicating that real-world dynamics frequently involve multiple interacting primitives. We also include 3,881 samples with no explicit interactions ($N = 0$). These scenes contain natural object motion without direct entity-to-entity interactions, providing a baseline for learning general video dynamics. Since **GraphVid** allows

Table 1: GraphVid-Bench statistics across ∼27K curated videos. Interaction complexity highlights the prevalence of multi-entity dynamics.

Metric	Mean	Med.	Std.
Nodes / Graph	7.76	5.0	8.03
Edges / Graph	4.85	4.0	3.46
Interaction Complexity			
No-Interaction (N=0)			3881
Single-Interaction (N=1)			10982
Multi-Interaction (N>1)			12641



(a) Taxonomy of interaction types. **(b)** Co-occurrence Statistics.

Fig. 4: Interaction statistics in GraphVid-Bench dataset. (a) Frequency of interaction primitives, grouped into four categories (force & manipulation, kinematics & motion, physical contact, and support & position). (b) Interaction co-occurrence matrix captures interaction primitives appearing together within the same video sequences. The co-occurrence of manipulation and motion primitives shows that real-world dynamics are inherently compositional, motivating the use of structured interaction graphs for controllable video generation.

users to add interaction edges at inference time, such samples help the model learn how scene behavior changes when new interactions are introduced. Appendix A provides additional details on the dataset curation process.

5 Experiments

5.1 Implementation Details

GraphVid leverages the pretrained LTX-Video [16] Diffusion Transformer as backbone for video generation. The DiT backbone and the VAE encoder–decoder are kept frozen during training. Object entities are extracted from the conditioning image using Qwen3-VL [2]. For each object, we obtain a $d_v = 8192$-dimensional visual embedding using mean and max pooling over object crops, a $d_t = 2560$-dimensional text embedding for the object label, and normalized bounding box coordinates. These are concatenated to form a $d_n = 10756$-dimensional node feature vector. Physical interactions between objects are encoded using Qwen3-Embedding [59], producing edge attributes of dimension $d_e = 2560$. During edge-aware graph reasoning, node and edge features are projected to a hidden dimension of $d = 512$, producing interaction-aware node embeddings. Node embeddings are projected to the LTX transformer hidden size using an MLP, mapping 512-dimensional graph embeddings to 4096 dimensions. Target videos are encoded into latent space using the pretrained VAE, and noise is injected at timesteps sampled from a logit-normal distribution. We employ LoRA modules of rank 128. During training, only the GNN parameters, graph-to-adapter projection layers, and LoRA weights are updated.

Table 2: Quantitative comparison on interaction-centric MoveBench subset. GraphVid achieves competitive performance across all metrics while using significantly fewer parameters and training data. Best and second-best are highlighted.

Method	Train Data	Trainable Params (B)	Inference (s)↓	FID↓	FVD↓	PSNR↑	SSIM↑	EPE↓
Wan-Move [8]	2M	14.5	1800	15.56	82.17	17.21	0.61	2.6
Motion-I2V [42]	10M	1.2	790	28.32	159.32	9.87	0.38	3.9
Tora [60]	630K	5	1200	24.45	110.47	11.27	0.54	3.3
MagicMotion [32]	23K	1.5	750	26.57	105.12	12.04	0.51	3.2
WISA [49]	80K	1	1000	25.98	107.89	15.08	0.53	4.1
FlashMotion [31]	23K	13	677	19.02	104.12	14.04	0.56	3.8
GraphVid (Ours)	27K	0.6	200	17.02	99.42	15.98	0.61	2.9

5.2 Experimental Setup

For evaluation, we establish a MoveBench [8] subset focused on interaction-centric scenarios as the primary evaluation suite. Since GraphVid emphasizes physics-aware relational control rather than explicit pixel-level motion supervision, we prioritize metrics that assess overall video quality, temporal coherence, and structural fidelity. In addition, we evaluate performance on a subset of the distilled MoveBench containing only multi-object interaction scenarios, allowing us to assess the effectiveness of GraphVid in more complex settings where reasoning over interactions between multiple entities is required. We report Frechet Video Distance (FVD) [46], Frechet Inception Distance (FID) [19], Peak Signal-to-Noise Ratio (PSNR), and Structural Similarity Index (SSIM) [51] to measure visual quality. We additionally report End-Point Error (EPE) [10], the average displacement between predicted and ground-truth optical flow, which directly measures whether the generated motion follows the intended interactions. Additional details can be found in Appendix B.

5.3 Quantitative Results

Table 2 compares GraphVid with existing controllable video generation methods on the interaction-centric MoveBench benchmark subset, where GraphVid achieves competitive performance across all metrics while requiring substantially fewer training samples, significantly fewer trainable parameters, and faster inference. Among prior approaches, WISA [49] is the most closely related method that explicitly incorporates structured physical knowledge into the video generation process. Compared to WISA, GraphVid achieves a substantial improvement in perceptual quality and reconstruction fidelity. In particular, GraphVid reduces FID from 25.98 to 17.02 and FVD from 107.89 to 99.42, corresponding to a 34.5% and 7.8% improvement, respectively, indicating significantly better visual realism and distribution alignment with ground-truth videos. Similarly, SSIM improves by 15.0%, and PSNR by 5.9%, demonstrating that our graph-based interaction modeling produces frames that are both sharper and structurally more consistent. This advantage also holds for motion accuracy, with GraphVid reducing EPE from 4.1 to 2.9, a 29.3% improvement over WISA, and achieving

Table 3: Multi-object quantitative comparison on interaction-centric MoveBench. GraphVid maintains strong structural coherence compared to larger baselines. Best and second-best are highlighted.

Method	Train Data	Trainable Params (B)	FID↓	FVD↓	PSNR↑	SSIM↑	EPE↓
Wan-Move [8]	2M	14.5	31.29	252	16.69	0.61	2.2
Tora [60]	630K	5	56.04	369	14.98	0.52	3.5
WISA [49]	80K	1	60.12	341	13.13	0.55	3.9
FlashMotion [31]	23K	13	55.03	311	12.12	0.52	3.9
GraphVid (Ours)	27K	0.6	49.45	291	14.44	0.55	3.0

the lowest EPE among all comparable-scale methods. Since EPE measures how closely the generated optical flow matches the ground truth, this indicates that our graph conditioning produces motion that most faithfully follows the specified interactions. The overall perceptual, reconstruction, and motion gains indicate that explicitly modeling interaction relationships through a directed scene graph provides a stronger inductive bias for controllable video generation than relying on predefined textual (physics-centric) attributes.

We further compare against trajectory-based control methods such as Motion-I2V [42], MagicMotion [32], FlashMotion [31], and Tora [60]. These approaches rely heavily on explicit motion signals such as trajectories, optical flow, or segmentation masks. Compared to Motion-I2V, GraphVid reduces FID by 39.9% (28.32 → 17.02) and FVD by 37.6% (159.32 → 99.42), while improving PSNR from 9.87 to 15.98 and SSIM from 0.38 to 0.61, indicating substantially better frame reconstruction and perceptual fidelity. Compared to FlashMotion, a recent trajectory-guided video generation method, GraphVid lowers FID by 10.5% (19.02 → 17.02) while also achieving stronger reconstruction quality, improving PSNR from 14.04 to 15.98 and SSIM from 0.56 to 0.61. Compared with MagicMotion, GraphVid reduces FID by approximately 36% (26.57 → 17.02) and improves temporal consistency. Notably, GraphVid attains a lower EPE (2.9) than all comparable trajectory-based methods (Motion-I2V 3.9, FlashMotion 3.8, Tora 3.3, MagicMotion 3.2), reducing motion error by 9.4–25.6% without relying on any explicit motion supervision. These improvements highlight the effectiveness of modeling scene dynamics through structured interaction reasoning. Wan-Move [8] achieves the best absolute metrics due to its significantly larger scale, being trained on approximately 2M videos and using a 14.5B parameter backbone. In contrast, GraphVid operates with only 0.6B trainable parameters and is trained on a much smaller 27K interaction-centric dataset. Despite using approximately two orders of magnitude less training data and ∼24× lower model params, GraphVid achieves competitive performance, outperforming most prior controllable methods on reconstruction and motion-accuracy.

A similar trend can be observed in Table 3, which evaluates performance on the multi-object subset of MoveBench where interaction reasoning becomes particularly important. In these scenarios, GraphVid improves multi-object interaction quality over other lightweight baselines, with clear relative gains, while using substantially fewer resources. When compared to models of similar data

scale, GraphVid achieves the best FID (49.45), improving over the next best (FlashMotion, 55.03) by 10.1%, and the best FVD (291), improving over the next best (FlashMotion, 311) by 6.4%; both indicating better perceptual realism and temporal consistency in multi-entity dynamics. GraphVid also achieves the lowest EPE among comparable-scale methods, confirming that its motion-accuracy advantage carries over to the harder multi-object setting. On reconstruction fidelity, GraphVid remains competitive with Tora (best PSNR) and WISA (best SSIM). Importantly, these gains come with far lower scale. GraphVid uses 0.6B trainable parameters (≈21.7× fewer than FlashMotion and 8.3× fewer than Tora) and trains on 27K clips (comparable to FlashMotion and 23× fewer than Tora), underscoring that interaction-graph conditioning provides strong multi-object coherence without requiring massive models or data. Appendix C presents additional results on our interaction-focused DAVIS [38] subset.

5.4 Efficiency and Scalability Analysis

Figure 5 plots inference throughput against model scale. Beyond accuracy, GraphVid offers strong efficiency advantages, achieving the fastest inference among all compared approaches at 200 seconds, significantly faster than Wan-Move (1800s), Tora (1200s), and WISA (1000s). This efficiency stems from our lightweight interaction graph conditioning and LoRA-based integration into the pre-trained DiT backbone, which avoids the need for heavy auxiliary motion encoders or large-scale motion prediction modules. This confirms that explicitly modeling interaction relationships through a structured scene graph provides a powerful and efficient alternative to trajectory-based motion supervision. By leveraging interaction-aware graph reasoning, GraphVid achieves strong controllability and physical plausibility while operating with substantially lower compute and data.

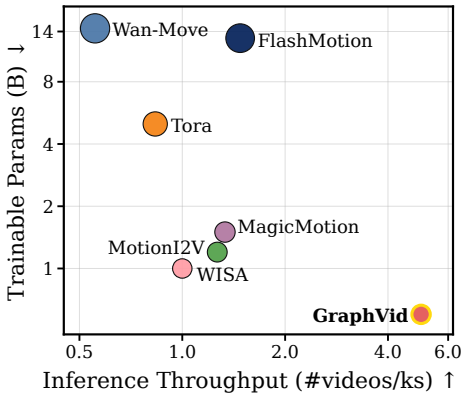


Fig. 5: Model efficiency comparison. Trade-off between inference throughput and scale. Bubble size reflects model scale.

5.5 Qualitative Analysis

Figure 6 presents representative qualitative results demonstrating that modeling scene dynamics through GraphVid interaction graphs enables precise and interpretable control over complex multi-entity motion. Examples illustrate diverse interaction types, including object–object interactions, human–object interactions, and articulated human motion. Across all scenarios, GraphVid preserves object identity and scene appearance, produces realistic motion trajectories

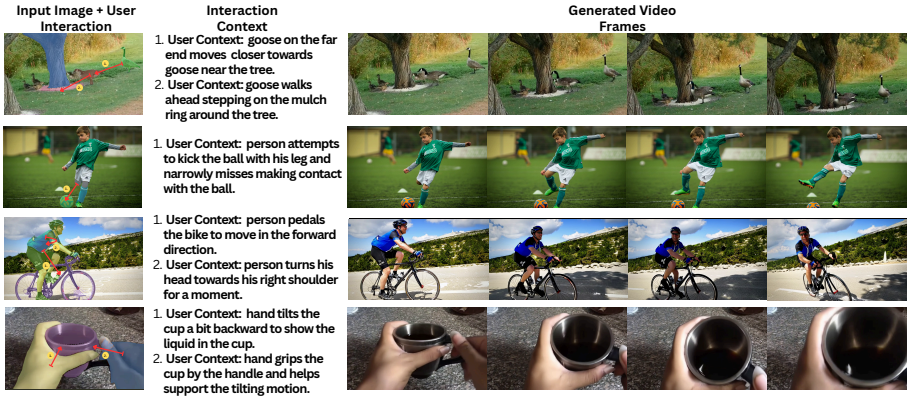


Fig. 6: Qualitative results of GraphVid for controllable interaction-based video generation. Given a single input image and user-specified interaction context, GraphVid generates temporally coherent video sequences reflecting the desired dynamics. Examples include object-object interactions (geese moving relative to each other), human-object interactions (kicking a ball, holding a cup), and articulated motion (cycling). The generated sequences demonstrate consistent motion, object preservation, and faithful execution of user-specified interaction dynamics.

and interaction-consistent behavior, and maintains stable background geometry. These examples indicate that the interaction-graph interface helps the model bind motion to the correct entities and combine multiple primitives into coherent multi-step dynamics, something that is typically brittle with purely text- or trajectory-only controls. Appendix D presents a human preference study, while Appendix E provides qualitative results across diverse interaction scenarios.

5.6 Ablation Studies

Backbone Generalization. To verify that relational scene graphs serve as a universal conditioning modality rather than an exploit of LTX's specific latent priors, we replace the 2B LTX backbone with the 5B Wan 2.2 model [48]. When training with this variant, our edge-aware GNN and adapter successfully map the topological graphs into Wan's latent space, achieving competitive performance (Table 4). This demonstrates that our graph representation is a backbone-agnostic conditioning signal capable of guiding scene dynamics across different foundation models.

Table 4: Backbone ablation.

Backbone	FID↓	FVD↓	PSNR↑	SSIM↑	EPE↓
Ours + LTX	**17.02**	**99.42**	**15.98**	**0.61**	**2.9**
Ours + Wan 2.2	17.29	100.01	15.70	0.60	3.1

Graph Sparsity and Attention Dilution. We evaluate the effect of the context window size on the DiT's self-attention mechanism by varying the maximum node capacity. As shown in Table 5, padding sparse real-world scene graphs to overly large

Table 5: Graph ablation.

#Nodes	FID↓	FVD↓	PSNR↑	SSIM↑
10	17.25	102.04	15.56	0.60
30	**17.02**	**99.42**	**15.98**	**0.61**
50	17.17	104.97	15.11	0.60
128	17.55	105.45	14.55	0.59
256	17.97	109.92	13.77	0.57

context windows can dilute attention and degrade generation quality. Capping the graph size at 30 nodes consistently outperforms larger capacities (128/256), achieving the best FVD (99.42). This suggests that a spatially tighter, denser node set yields better structural alignment: as the maximum node budget increases (*e.g.*, from 30 → 256), FID degrades, likely because the additional capacity is dominated by zero-padding rather than informative structure. The resulting padded tokens inject noise into attention and dilute meaningful relational cues, making it harder for the transformer to concentrate on active interactions. Overall, these results suggest that balanced, densely populated graph representations provide the strongest conditioning signal, and GraphVid therefore uses a maximum node capacity of 30 for all experiments.

Conditioning Modalities and Edge Semantics. To evaluate the importance of structured conditioning and edge semantics, we perform an inference-time ablation comparing three distinct regimes. Table 6 shows (1) the standard LTX baseline, which generates video using a global text prompt and an initial image; (2) GraphVid conditioned on the image and a topology-only graph, where the user provides interaction arrows and overlays but omits textual edge context; and (3) our full GraphVid approach, utilizing the image and a complete graph equipped with explicit text context for each interaction edge. While the text-based baseline achieves moderate visual quality, it struggles to bind localized motion to specific entities (FID 20.04). Replacing the global text with our topology-only graph improves FID to 17.52, indicating that explicitly graphing spatial boundaries aids motion localization. However, without the semantic richness of the textual interaction embeddings, the model exhibits severe motion ambiguity, failing to distinguish opposing interactions like "pushing" and "pulling". The full graph conditioning outperforms both baselines, reducing FID to 17.02 and FVD to 99.42. These results suggest that graph topology alone provides useful spatial grounding, while semantic edge labels supply additional interaction intent that helps disambiguate object relationships and improve motion generation.

Table 6: Conditioning modalities.

Variant	FID↓	FVD↓	PSNR↑	SSIM↑	EPE↓
Text + Img (LTX)	20.04	109.34	11.23	0.53	3.6
Graph (no edge text)	17.52	101.32	**16.88**	0.60	3.0
Graph (full)	**17.02**	**99.42**	15.98	**0.61**	**2.9**

LoRA Rank and Capacity Trade-offs. We inject low-rank adaptation (LoRA) matrices into the transformer blocks to modulate the pretrained weights with our graph conditions. In this ablation, we study the effect of the LoRA rank on generation quality and reconstruction fidelity. As shown in Table 7, increasing the rank consistently improves performance across all metrics. Moving from rank 16 to 128 reduces FID from 18.34 to 17.02 (–7.2%) and FVD from 103.18 to 99.42 (–3.6%), indicating better perceptual quality and temporal coherence. At the same time, recon-

Table 7: Ablation on LoRA rank.

Rank	FID↓	FVD↓	PSNR↑	SSIM↑
16	18.34	103.18	13.70	0.58
32	17.89	102.39	14.51	0.60
64	17.53	100.99	15.25	0.60
128	**17.02**	**99.42**	**15.98**	**0.61**

struction metrics improve substantially, with PSNR increasing from 13.70 to 15.98 (+16.6%) and SSIM from 0.58 to 0.61 (+5.2%). These results suggest that higher-rank adapters provide greater representational capacity to capture complex motion and appearance variations, leading to more faithful and temporally consistent video generation. Rank 128 is used in all the reported experiments.

6 Conclusion

We introduce **GraphVid**, a framework for controllable image-to-video generation that represents scene dynamics through directed interaction graphs. Unlike prior approaches that rely on dense motion signals or low-level physics annotations, GraphVid models motion as structured interactions between entities, enabling intuitive multi-object control directly from a single image. By integrating Edge-Aware Graph Reasoning with a frozen video diffusion transformer through parameter-efficient LoRA conditioning, GraphVid preserves strong pretrained generative priors while enabling structured control over scene dynamics. To support this paradigm, we also curate **GraphVid-Bench**, a dataset of 27K interaction-centric videos paired with explicit interaction graphs, providing supervision for learning relational video dynamics. Experiments show that GraphVid matches or outperforms prior controllable video generation methods while using far fewer parameters and training data, highlighting interaction graphs as a scalable interface for controllable video generation.

Acknowledgments

This research was partially supported by Google, the Google TPU Research Cloud (TRC) program, the NSF CAREER Award #2542328, the U.S. Defense Advanced Research Projects Agency (DARPA) under award HR001125C0303, and the U.S. Army under contract W5170125CA160. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of Google, NSF, DARPA, the U.S. Army, or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for governmental purposes notwithstanding any copyright annotation therein. This work also used the Delta GPUs provided by the National Center for Supercomputing Applications through allocations CIS240808 and CIS250318 from the Advanced Cyberinfrastructure Coordination Ecosystem: Services & Support (ACCESS [6]) program, supported by National Science Foundation grants #2138259, #2138286, #2138307, #2137603, and #2138296.

References

1. Armeni, I., He, Z.Y., Gwak, J., Zamir, A.R., Fischer, M., Malik, J., Savarese, S.: 3d scene graph: A structure for unified semantics, 3d space, and camera. In: IEEE Conf. Comput. Vis. Pattern Recog. (2019)

2. Bai, S., Cai, Y., Chen, R., Chen, K., Chen, X., Cheng, Z., Deng, L., Ding, W., Gao, C., Ge, C., et al.: Qwen3-vl technical report. arXiv preprint arXiv:2511.21631 (2025)
3. Bar, A., Herzig, R., Wang, X., Rohrbach, A., Chechik, G., Darrell, T., Globerson, A.: Compositional video synthesis with action graphs. In: Int. Conf. Mach. Learn. (2021)
4. Bar-Tal, O., Chefer, H., Tov, O., Herrmann, C., Paiss, R., Zada, S., Ephrat, A., Hur, J., Liu, G., Raj, A., et al.: Lumiere: A space-time diffusion model for video generation. In: SIGGRAPH Asia Conference Papers (2024)
5. Blattmann, A., Dockhorn, T., Kulal, S., Mendelevitch, D., Kilian, M., Lorenz, D., Levi, Y., English, Z., Voleti, V., Letts, A., et al.: Stable video diffusion: Scaling latent video diffusion models to large datasets. arXiv preprint arXiv:2311.15127 (2023)
6. Boerner, T.J., Deems, S., Furlani, T.R., Knuth, S.L., Towns, J.: ACCESS: Advancing Innovation: NSF's Advanced Cyberinfrastructure Coordination Ecosystem: Services & Support. In: Practice and Experience in Advanced Research Computing (PEARC) (2023)
7. Chang, X., Ren, P., Xu, P., Li, Z., Chen, X., Hauptmann, A.: A comprehensive survey of scene graphs: Generation and application. IEEE Transactions on Pattern Analysis and Machine Intelligence **45**(1), 1–26 (2021)
8. Chu, R., He, Y., Chen, Z., Zhang, S., Xu, X., Xia, B., WANG, D., Yi, H., Liu, X., Zhao, H., et al.: Wan-move: Motion-controllable video generation via latent trajectory guidance. In: Adv. Neural Inform. Process. Syst. (2025)
9. Dutta, A., Mehrab, K.S., Sawhney, M., Neog, A., Khurana, M., Fatemi, S., Pradhan, A., Maruf, M., Lourentzou, I., Daw, A., et al.: Open world scene graph generation using vision language models. arXiv preprint arXiv:2506.08189 (2025)
10. Geng, D., Herrmann, C., Hur, J., Cole, F., Zhang, S., Pfaff, T., Lopez-Guevara, T., Aytar, Y., Rubinstein, M., Sun, C., et al.: Motion prompting: Controlling video generation with motion trajectories. In: IEEE Conf. Comput. Vis. Pattern Recog. (2025)
11. Geyer, M., Bar-Tal, O., Bagon, S., Dekel, T.: Tokenflow: Consistent diffusion features for consistent video editing. arXiv preprint arXiv:2307.10373 (2023)
12. Gillman, N., Herrmann, C., Freeman, M., Aggarwal, D., Luo, E., Sun, D., Sun, C.: Force prompting: Video generation models can learn and generalize physics-based control signals. In: Adv. Neural Inform. Process. Syst. (2025)
13. Goyal, R., Ebrahimi Kahou, S., Michalski, V., Materzynska, J., Westphal, S., Kim, H., Haenel, V., Fruend, I., Yianilos, P., Mueller-Freitag, M., et al.: The" something something" video database for learning and evaluating visual common sense. In: IEEE Conf. Comput. Vis. Pattern Recog. (2017)
14. Guo, Y., Yang, C., Rao, A., Agrawala, M., Lin, D., Dai, B.: Sparsectrl: Adding sparse controls to text-to-video diffusion models. In: Eur. Conf. Comput. Vis. (2024)
15. Guo, Y., Yang, C., Rao, A., Liang, Z., Wang, Y., Qiao, Y., Agrawala, M., Lin, D., Dai, B.: Animatediff: Animate your personalized text-to-image diffusion models without specific tuning. Int. Conf. Learn. Represent. (2024)
16. HaCohen, Y., Chiprut, N., Brazowski, B., Shalem, D., Moshe, D., Richardson, E., Levin, E., Shiran, G., Zabari, N., Gordon, O., et al.: Ltx-video: Realtime video latent diffusion. arXiv preprint arXiv:2501.00103 (2024)
17. Hamilton, W., Ying, Z., Leskovec, J.: Inductive representation learning on large graphs. Adv. Neural Inform. Process. Syst. (2017)

18. He, Y., Liu, Z., Chen, J., Tian, Z., Liu, H., Chi, X., Liu, R., Yuan, R., Xing, Y., Wang, W., et al.: Llms meet multimodal generation and editing: A survey. arXiv preprint arXiv:2405.19334 (2024)
19. Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Hochreiter, S.: Gans trained by a two time-scale update rule converge to a local nash equilibrium. Adv. Neural Inform. Process. Syst. (2017)
20. Holla, M., Lourentzou, I.: Commonsense for zero-shot natural language video localization. In: AAAI (2024)
21. Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W., et al.: Lora: Low-rank adaptation of large language models. Int. Conf. Learn. Represent. (2022)
22. Hu, Y., Chen, Z., Luo, C.: Lamd: Latent motion diffusion for image-conditional video generation. Int. J. Comput. Vis. (2025)
23. Huang, Z., He, Y., Yu, J., Zhang, F., Si, C., Jiang, Y., Zhang, Y., Wu, T., Jin, Q., Chanpaisit, N., et al.: Vbench: Comprehensive benchmark suite for video generative models. In: IEEE Conf. Comput. Vis. Pattern Recog. (2024)
24. Karaev, N., Rocco, I., Graham, B., Neverova, N., Vedaldi, A., Rupprecht, C.: Cotracker: It is better to track together. In: Eur. Conf. Comput. Vis. (2024)
25. Karwande, G., Mbakwe, A.B., Wu, J.T., Celi, L.A., Moradi, M., Lourentzou, I.: Chexrelnet: An anatomy-aware model for tracking longitudinal relationships between chest x-rays. In: International Conference on Medical Image Computing and Computer-Assisted Intervention (2022)
26. Khan, M.J., Ilievski, F., Breslin, J.G., Curry, E.: A survey of neurosymbolic visual reasoning with scene graphs and common sense knowledge. Neurosymbolic Artificial Intelligence (2025)
27. Kipf, T.N., Welling, M.: Semi-supervised classification with graph convolutional networks. arXiv preprint arXiv:1609.02907 (2016)
28. Ku, M., Wei, C., Ren, W., Yang, H., Chen, W.: Anyv2v: A tuning-free framework for any video-to-video editing tasks. arXiv preprint arXiv:2403.14468 (2024)
29. Lei, G., Wang, C., Zhang, R., Wang, Y., Li, H., Xu, W.: Animateanything: Consistent and controllable animation for video generation. In: IEEE Conf. Comput. Vis. Pattern Recog. (2025)
30. Li, H., Zhu, G., Zhang, L., Jiang, Y., Dang, Y., Hou, H., Shen, P., Zhao, X., Shah, S.A.A., Bennamoun, M.: Scene graph generation: A comprehensive survey. Neurocomputing (2024)
31. Li, Q., Xing, Z., Wang, R., Cao, H., Dai, Q., Dong, D., Wu, Z.: Flashmotion: Few-step controllable video generation with trajectory guidance. In: IEEE Conf. Comput. Vis. Pattern Recog. (2026)
32. Li, Q., Xing, Z., Wang, R., Zhang, H., Dai, Q., Wu, Z.: Magicmotion: Controllable video generation with dense-to-sparse trajectory guidance. In: IEEE Conf. Comput. Vis. Pattern Recog. (2025)
33. Li, X., Ma, C., Yang, X., Yang, M.H.: Vidtome: Video token merging for zero-shot video editing. In: IEEE Conf. Comput. Vis. Pattern Recog. (2024)
34. Li, Y., Wang, X., Zhang, Z., Wang, Z., Yuan, Z., Xie, L., Shan, Y., Zou, Y.: Image conductor: Precision control for interactive video synthesis. In: AAAI (2025)
35. Lin, B., Ge, Y., Cheng, X., Li, Z., Zhu, B., Wang, S., He, X., Ye, Y., Yuan, S., Chen, L., et al.: Open-sora plan: Open-source large video generation model. arXiv preprint arXiv:2412.00131 (2024)
36. Ogunleye, M.A., Abdelrahman, E., Lourentzou, I.: 3d-vcd: Hallucination mitigation in 3d-llm embodied agents through visual contrastive decoding. In: IEEE Conf. Comput. Vis. Pattern Recog. (2026)

37. Peebles, W., Xie, S.: Scalable diffusion models with transformers. In: IEEE Conf. Comput. Vis. Pattern Recog. (2023)
38. Pont-Tuset, J., Perazzi, F., Caelles, S., Arbeláez, P., Sorkine-Hornung, A., Van Gool, L.: The 2017 davis challenge on video object segmentation. arXiv preprint arXiv:1704.00675 (2017)
39. Romero, D., Bermudez, A., Li, H., Pizzati, F., Laptev, I.: Learning to generate rigid body interactions with video diffusion models. arXiv preprint arXiv:2510.02284 (2025)
40. Shen, Y., Liu, J., Li, X., Liu, Y., Li, B., Yang, H., Jia, W., Li, Y., Yu, T., Rehg, J.M., Cao, X., Lourentzou, I.: Egoforge: Goal-directed egocentric world simulator. arXiv preprint arXiv:2603.20169 (2026)
41. Shen, Y., Xiong, J., Yu, T., Lourentzou, I.: Phantom: Physics-infused video generation via joint modeling of visual and latent physical dynamics. In: IEEE Conf. Comput. Vis. Pattern Recog. (2026)
42. Shi, X., Huang, Z., Wang, F.Y., Bian, W., Li, D., Zhang, Y., Zhang, M., Cheung, K.C., See, S., Qin, H., et al.: Motion-i2v: Consistent and controllable image-to-video generation with explicit motion modeling. In: SIGGRAPH Conference Papers (2024)
43. Singer, U., Polyak, A., Hayes, T., Yin, X., An, J., Zhang, S., Hu, Q., Yang, H., Ashual, O., Gafni, O., et al.: Make-a-video: Text-to-video generation without text-video data. arXiv preprint arXiv:2209.14792 (2022)
44. Susladkar, O., Prakash, T., Juvekar, A., Nguyen, K.A., Jang, D.H., Dhillon, I.S., Lourentzou, I.: Pyratok: Language-aligned pyramidal tokenizer for video understanding and generation. In: IEEE Conf. Comput. Vis. Pattern Recog. (2026)
45. Susladkar, O., Sen Gupta, J., Sehgal, C., Mittal, S., Singhal, R.: Motionaura: Generating high-quality and motion consistent videos using discrete diffusion. In: Int. Conf. Learn. Represent. (2025)
46. Unterthiner, T., Van Steenkiste, S., Kurach, K., Marinier, R., Michalski, M., Gelly, S.: Fvd: A new metric for video generation. Int. Conf. Learn. Represent. Worksh. DeepGenStruct (2019)
47. Wald, J., Dhamo, H., Navab, N., Tombari, F.: Learning 3d semantic scene graphs from 3d indoor reconstructions. In: IEEE Conf. Comput. Vis. Pattern Recog. (2020)
48. Wan, T., Wang, A., Ai, B., Wen, B., Mao, C., Xie, C.W., Chen, D., Yu, F., Zhao, H., Yang, J., et al.: Wan: Open and advanced large-scale video generative models. arXiv preprint arXiv:2503.20314 (2025)
49. Wang, J., Ma, A., Cao, K., Zheng, J., Feng, J., Zhang, Z., Pang, W., Liang, X.: Wisa: World simulator assistant for physics-aware text-to-video generation. In: Adv. Neural Inform. Process. Syst. (2025)
50. Wang, X., Yuan, H., Zhang, S., Chen, D., Wang, J., Zhang, Y., Shen, Y., Zhao, D., Zhou, J.: Videocomposer: Compositional video synthesis with motion controllability. Adv. Neural Inform. Process. Syst. (2023)
51. Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P.: Image quality assessment: from error visibility to structural similarity. IEEE Int. Conf. Image Process. (2004)
52. Wei, Y., Zhang, S., Yuan, H., Gong, B., Tang, L., Wang, X., Qiu, H., Li, H., Tan, S., Zhang, Y., et al.: Dreamrelation: Relation-centric video customization. In: IEEE Conf. Comput. Vis. Pattern Recog. (2025)
53. Wu, W., Li, Z., Gu, Y., Zhao, R., He, Y., Zhang, D.J., Shou, M.Z., Li, Y., Gao, T., Zhang, D.: Draganything: Motion control for anything using entity representation. In: Eur. Conf. Comput. Vis. (2024)

54. Xu, K., Hu, W., Leskovec, J., Jegelka, S.: How powerful are graph neural networks? arXiv preprint arXiv:1810.00826 (2018)
55. Yang, X., He, C., Ma, J., Zhang, L.: Motion-guided latent diffusion for temporally consistent real-world video super-resolution. In: Eur. Conf. Comput. Vis. (2024)
56. Yang, Z., Teng, J., Zheng, W., Ding, M., Huang, S., Xu, J., Yang, Y., Hong, W., Zhang, X., Feng, G., et al.: Cogvideox: Text-to-video diffusion models with an expert transformer. In: Int. Conf. Learn. Represent. (2025)
57. Yin, S., Wu, C., Liang, J., Shi, J., Li, H., Ming, G., Duan, N.: Dragnuwa: Fine-grained control in video generation by integrating text, image, and trajectory. arXiv preprint arXiv:2308.08089 (2023)
58. Zhang, L., Rao, A., Agrawala, M.: Adding conditional control to text-to-image diffusion models. In: IEEE Conf. Comput. Vis. Pattern Recog. (2023)
59. Zhang, Y., Li, M., Long, D., Zhang, X., Lin, H., Yang, B., Xie, P., Yang, A., Liu, D., Lin, J., et al.: Qwen3 embedding: Advancing text embedding and reranking through foundation models. arXiv preprint arXiv:2506.05176 (2025)
60. Zhang, Z., Liao, J., Li, M., Dai, Z., Qiu, B., Zhu, S., Qin, L., Wang, W.: Tora: Trajectory-oriented diffusion transformer for video generation. In: IEEE Conf. Comput. Vis. Pattern Recog. (2025)